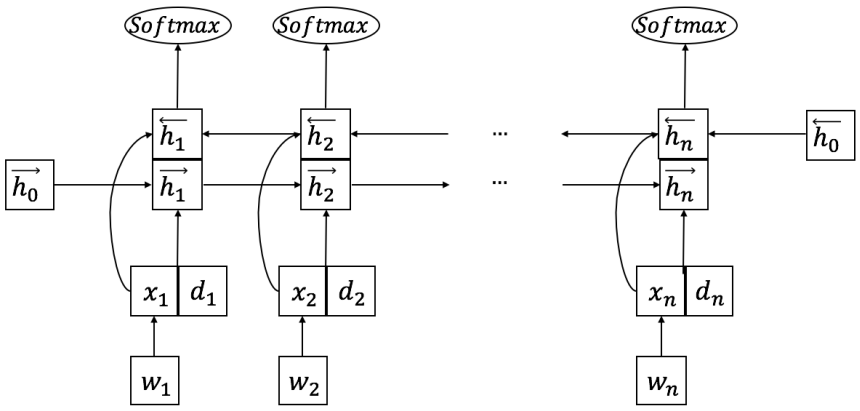
DS-GA 1006: Roivant Sciences Drug Names Identification

2nd Status Update

Kailing Wang (kailing.wang@nyu.edu)

➢ Model

In the past two weeks, we built a many-to-many RNN model. This additional model is composed of three major components: word-embedding, bi-directional RNN, and the top-layer classifier. A sentence is indicated as $w = (w_1, \ldots, w_n)$, where n is the length of the sentence. We concatenate the word-embedding with its other information, including its POS tagging, dependency relation and "Target" indicator. To represent words and their context in a sentence, these pairs are fed into a bi-directional RNN to obtain hidden vectors $h_t = [\overrightarrow{h_t}; \overleftarrow{h_t}]$. Finally, the hidden vectors are fed into the Softmax function to determine whether each of words is a drug name (1), target name (2), or neither (0).



➢ Result

	learning_rate	vocab_size	emb_dim	hidden_size	num_epochs	RNN_type	num_layers	add_info_into_RNN	Loss	train_recal	val_recall	test_recall	test_precision	F-measure
model_1	0.001	20000	50	100	50	LSTM	1	1	4.95	98.98%	68.27%	59.45%	83.68%	0.695141
model_2	0.001	20000	50	100	50	LSTM	1	0	1.97E-05	98.38%	63.05%	56.82%	84.68%	0.680073
model_3	0.001	20000	50	100	50	GRU	1	1	1.40E-06	98.88%	70.28%	60.93%	87.87%	0.719613
model_4	0.001	20000	50	100	50	GRU	1	0	2.73E-06	99.39%	74.30%	65.42%	79.73%	0.718696
model_5	0.001	6000	50	100	50	GRU	1	0	6.31E-06	99.19%	69.48%	58.88%	83.78%	0.69157
model_6	0.001	6000	50	100	50	LSTM	1	1	5.47E-10	98.68%	73.09%	62.80%	77.24%	0.692755
model_7	0.001	20000	50	100	50	GRU	2	1	2.64E-06	99.80%	70.68%	59.81%	86.72%	0.70794

➢ Discussion

As the labels of this many-to-many model are generated based on ChEMBL dataset, which means that it is assumed that there is only one noun phrase labeled as "Drug" in each sentence. However, the fact is that there could be multiple drug names appear in a single sentence. So therefore, in the next step, we will improve the labeling method, and then re-run the model (to see whether we get better performance). Once this many-to-many NER model is desirable, we will incorporate it with our previous attention-based classification model.

Still, we could try to replace the learned word-embedding with a pre-trained one (e.g. Glove).